

A scalable and efficient covariate selection criterion for mixed effects regression models with unknown random effects structure

Radu V. Craiu

*Department of Statistics
University of Toronto
Toronto, Ontario, M5S 3G3, Canada
e-mail: craiu@utstat.utoronto.ca*

Thierry Duchesne

*Département de mathématiques et de statistique
Université Laval
Québec, Québec, G1V 0A6, Canada
e-mail: thierry.duchesne@mat.ulaval.ca*

Abstract:

We propose a new model selection criterion for mixed effects regression models that is computable even when the structure and the distribution of the random effects are unknown. The criterion is most useful in the early stage of the model building process when one needs to decide which covariates should be included in a mixed effects regression model but has no knowledge of the random effect structure. The calculation of the criterion requires only the evaluation of cluster-level log-likelihoods and does not rely on heavy numerical integration. We provide theoretical and numerical arguments to justify the method and we illustrate its usefulness by analyzing data on the consumption of alcohol by young American Indians.

AMS 2000 subject classifications: Primary 62F07; secondary 62J12.

Keywords and phrases: Akaike Information Criterion, Generalized linear mixed model, h-likelihood, Random coefficient model, Two-stage estimation, Variable selection.

1. Introduction

Studies where a large number of observations are collected for each experimental unit, or cluster, are quite common. For instance, in behavioural ecology animals that wear GPS collars are tracked and data for each individual are collected every hour for months or years; in marketing studies banks record every credit card transaction made by a client; in some epidemiological studies data are collected on physicians who each treat a large number of patients; in criminology, data are recorded at every contact of a repeat offender with the justice system. In the social study that we use to illustrate our method, a large number of students are surveyed in a number of secondary high schools. In many instances where such data are collected, analysts will account for the dependence within each cluster by fitting a mixed effects regression model. In the construction of the latter an important and early step concerns selecting the covariates that are included in the model.

The importance of variable selection has been recognized in statistics and there is a vast body of work devoted to developing criteria for this problem (e.g., see the book of [Burnham and Anderson, 2002](#)). Traditionally, the Akaike information criterion (AIC) introduced in the foundational work of [Akaike \(1970\)](#) along with its small sample corrections ([Hurvich and Tsai, 1989](#); [Cavanaugh, 1997](#)), and the Bayesian Information Criterion (BIC), introduced by [Schwarz \(1978\)](#), have been among the first methods used to select the covariates in regression models with fixed effects. All these are special cases of the Generalized Information Criterion (GIC) ([Nishii, 1984](#); [Shibata, 2005](#); [Rao and Wu, 1989](#)) where the aim is to find the model M that minimizes

$$-\mathcal{L}(M) + \lambda|M|, \quad (1)$$

where $\mathcal{L}(M)$ is a measure of fit and $\lambda|M|$ is the penalty incurred by a model with size $|M|$. The GIC proposed by [Rao and Wu \(1989\)](#) is a strongly consistent variable selection criterion with a flexible penalty function. Other criteria with data adaptive penalty functions are obtained from a different theoretical paradigm that is based on the communication theory developed by Claude Shannon ([Shannon, 1948](#)). The Minimum Description Length (MDL) principle of [Rissanen \(1989\)](#) will prefer the model that yields the shortest code length required to describe the data. MDL's applications to statistical problems have shown promising results (see, for instance, [Barron et al., 1998](#); [Quinlan and Rivest, 1989](#); [Lee, 2000, 2001](#); [Hansen and Yu, 2001](#); [Craiu and Lee, 2005](#); [Hansen and Yu, 2003](#); [Li et al., 2014](#)).

The introduction of mixed effects models required new strategies for selecting both the fixed and the random effects. In this context, whether the inferential focus is on marginal or conditional model parameters becomes relevant as these two scenarios require separate treatments. While in the former case one could use the traditional criteria to select the covariates in the model, the latter considers the choice of covariates conditional on random effects. In [Vaida and Blanchard \(2005\)](#) the authors proposed the conditional AIC (cAIC) for situations in which the inferential focus is on cluster-specific parameters. Subsequently, the cAIC for linear mixed models has been further expanded by [Liang and Wu \(2008\)](#), [Greven and Kneib \(2010\)](#) and [Saefken et al. \(2014\)](#) who account for the estimation of variance parameters and by [Donohue et al. \(2011\)](#) who have extended cAIC to generalized linear mixed models (GLMM) and survival models with random effects. [Yu et al. \(2013\)](#) have proposed a further adjustment for cAIC in GLMM when the variance components must be estimated. An alternative BIC suitable for mixed effects models has been proposed by [Delattre et al. \(2014\)](#). In a departure from classical approaches, [Jiang et al. \(2008\)](#) propose a method in which incorrect models are fenced off and the best model is selected from the remaining ones. An excellent review of the methods briefly discussed here and others can be found in [Müller et al. \(2013\)](#).

Our current contribution for a new criterion is motivated by GLMM applications in ecology and social sciences where model selection is traditionally based on information criteria and where (i) little is known about the structure of the random effects *a priori* and (ii) numerical approximations of the marginal likelihood may be challenging due to model and data size ([Craiu et al., 2011](#); [Molenberghs et al., 2011](#)). The new criterion is intended as a first covariate filter in the early stage of the analysis. Given this aim, it is important that the proposed criterion is computable without the need to specify the random effect structure and that one does not incur a heavy computational cost in order to calculate its value for the submodels under consideration. After this initial stage, other numerically demanding methods can be exploited to search in the smaller model space.

The key strategy in the derivation of this new criterion is to avoid computation of the marginal log-likelihood function or global maximum likelihood estimators, which requires the specification of a random effect structure and the use of numerical integration. The criterion developed here is suitable for “partitioned data” methods (sometimes referred to as “divide-and-conquer” approaches) that have been proposed to fit mixed effects models when the

data are large or have a complex structure. Such methods include the two-stage approach of Korn and Whittemore (1979) and Stiratelli et al. (1984), the CREML method of Chervoneva et al. (2006), the two-step method of Craiu et al. (2011) or the pseudo-likelihood approach of Molenberghs et al. (2011). All these methods have in common that they fit separate simple models to each element of a partition of the data and then suitably unify the analyses for these simple models to produce inference for the global mixed effects model. Since these approaches do not resort to numerical integration, they make it easier to avoid computational issues and they can even remain asymptotically efficient in some specific cases.

In this paper we focus on deriving a criterion for filtering the potential covariates for use in a standard GLMM as described, for instance, in Chapter 3 of Jiang (2007). The proposed criterion, called meanAIC, is easy to compute and it does not require the specification of the random effects structure. We give a theoretical development of meanAIC along with heuristic arguments that justify it. Our simulation study shows that the proposed criterion exhibits good finite sample performance.

The remainder of the paper is organized as follows. Section 2 presents the data and model. The new criterion is developed and justified in Section 3. The simulation study is presented in Section 4 and a real data illustration forms Section 5. The paper concludes with a discussion and ideas for future work.

2. Data and model

2.1. Population and data

We consider a population of independent clusters, each containing a number of individual observations of the form (Y, x_1, \dots, x_p) with Y being a response variable and x_1, \dots, x_p potential explanatory variables. We assume that the distribution of Y given x_1, \dots, x_p is given by a generalized linear model whose regression coefficients may vary from cluster to cluster. The statistical model described below will assume that all the responses in the same cluster share some commonality that makes them dependent. We assume that there are K clusters and n_i data points in each cluster, $1 \leq i \leq K$.

2.2. Generalized linear mixed model (GLMM)

Let $Y_i = (Y_{i1}, \dots, Y_{in_i})^\top$ be the response vector for cluster i and $X_i = (x_{0i}, x_{i1}, \dots, x_{ir})$ be the corresponding covariate matrix, with $x_{ik} = (x_{i1k}, \dots, x_{in_ik})^\top$, $k = 0, \dots, r$ and x_{0i} an n_i -vector with all entries equal to 1. Throughout the paper the value of the k th covariate for the j th individual in the i th cluster will be denoted x_{ijk} . The context will clarify whether x_{ij} refers to the j th row of X_i (of length r) or x_{ik} refers to the k th column of X_i (of length n_i).

The dependence among observations in cluster i will be captured using the random vector b_i , where $\{b_i \in \mathbf{R}^q : i = 1, \dots, K\}$ are assumed to be iid with cumulative distribution function (cdf) H and probability density function (pdf) h . Throughout the paper we suppose that $q \leq r$. Let \mathcal{J} be a subset of size s of $\{0, \dots, q\}$, $Z_i = \{x_{ik}, k \in \mathcal{J}\}$ and $\beta = (\beta_0, \dots, \beta_r)^\top$. Under our population assumption and sampling scheme: i) (Y_i, X_i) , $i = 1, \dots, K$, are independent and ii) for all $1 \leq i \leq K$ $\{Y_{ij} : 1 \leq j \leq n_i\}$ are independent given X_i and b_i , iii) the distribution of $Y_{ij}|b_i, X_i$ belongs to the exponential family with pdf f_{ij} and

$$\mu_{ij} = E[Y_{ij}|b_i, X_i] = g^{-1}(\beta^\top x_{ij} + b_i^\top z_{ij}), \quad (2)$$

where x_{ij}^\top and z_{ij}^\top denote the j -th row of X_i and Z_i , respectively, and g is a known link function. This is the usual GLMM with random regression coefficients (Ch. 3 in Jiang, 2007). Let $\tilde{x}_{ij} = x_{ij} + \tilde{z}_{ij}$ where $\tilde{z}_{ijk} = z_{ijk}$ if $k \in \mathcal{J}$ and 0 otherwise. Similarly, let $\tilde{\beta}_i = \beta_i + \tilde{b}_i$ with $\tilde{b}_{ik} = b_{ik}$ if $k \in \mathcal{J}$ and 0 otherwise. Then (2) can be written as

$$\mu_{ij} = E[Y_{ij}|b_i, X_i] = g^{-1}(\tilde{\beta}_i^\top \tilde{x}_{ij}) \quad (3)$$

and the average conditional log-likelihood contribution in cluster i can be written as

$$\bar{\ell}_{n_i}(\tilde{\beta}_i) = n_i^{-1} \sum_{j=1}^{n_i} \log f_{ij}(Y_{ij}; \tilde{\beta}_i). \quad (4)$$

When one wishes to keep all r covariates in the model, H is the multivariate normal with mean 0 and variance matrix D and the subset \mathcal{J} and the structure of the matrix D are known, then inference methods for β and D , as well as predictions for b_i , are widely available in standard software. They are typically based on standard maximum likelihood, residual maximum likelihood, penalized quasi-likelihood or Bayesian methods. However, in many

practical situations all r covariates are not required and the random effect structure (subset \mathcal{J}) and distribution (H) are not known. In the following section we will derive a criterion that can guide the selection of covariates without having to specify neither \mathcal{J} nor H .

3. The meanAIC criterion

When fitting a mixed effects model, [Vaida and Blanchard \(2005\)](#) and [Greven and Kneib \(2010\)](#) argue that the selection of fixed effects and other marginal population parameters can be performed using the marginal Akaike information criterion (henceforth denoted mAIC). However, mAIC is based on the maximized marginal log-likelihood, whose computation involves marginalizing out the random effects via numerical integration. A challenge is that the latter may become numerically cumbersome when the dimension s of the random effects is large or when the data are massive. An even more serious problem is that marginal likelihood calculation requires the specification of the random effect structure \mathcal{J} and distribution H .

When considering a partitioned data approach, one realizes that the GLMM specification yields ordinary independent GLMs in each cluster. Many two-stage estimation approaches (e.g., [Stiratelli et al., 1984](#); [Renard et al., 2004](#); [Chervoneva et al., 2006](#); [Craiu et al., 2011](#); [Molenberghs et al., 2011](#)) rely on (some of) the following cluster-specific information that is usually easier to obtain than their full data counterparts:

- n_i , the number of observations in cluster i ;
- $\hat{\beta}_i = \arg \max_{\beta} \bar{\ell}_{n_i}(\beta)$, the MLE of β in cluster i ;
- $\bar{\ell}_{n_i}(\hat{\beta}_i)$, the maximized average log-likelihood in cluster i ;
- $H_i = \frac{\partial^2}{\partial \beta \partial \beta^\top} \bar{\ell}_{n_i}(\beta) \Big|_{\beta=\hat{\beta}_i}$, the Hessian of the average log-likelihood evaluated at $\hat{\beta}_i$.

It is clear that a criterion that would be based only on these four elements from each cluster should be easy to compute in practice. When the number of clusters is large, computations can be easily parallelized with one CPU fitting an ordinary GLM to its assigned cluster and returning the required cluster-level objects.

3.1. Derivation of meanAIC

Without specifying the random effects distribution one cannot compute the marginal likelihood that is used for deriving the marginal AIC (mAIC) criterion. Instead, we consider the h-likelihood proposed by [Lee and Nelder \(1996\)](#) which they introduced as the “natural generalization of the Fisher likelihood” for models with random effects. We assume that the true data generating model is of the form $p_0(y, b) = f_0(y|b)h_0(b)$. A candidate model has h-likelihood given by $p(y, b; \beta) = f(y|b; \beta)h(b)$. The conditional pdf f of y given the random effects b belongs to the exponential class modelled via a GLMM with coefficients $\beta + b$. The Kulback-Leibler divergence between p_0 and p is then

$$\begin{aligned} \Delta(p, p_0) &= E_{f_0, h_0} \log \left\{ \frac{f_0(y|b)h_0(b)}{f(y|b; \beta)h(b)} \right\} \\ &= \int \int \log \{f_0(y|b)h_0(b)\} f_0(y|b)h_0(b) dy db \end{aligned} \quad (5)$$

$$- \int \int f_0(y|b)h_0(b) \log h(b) dy db \quad (6)$$

$$- \int h_0(b) \left\{ \int \log f(y|b; \beta) f_0(y|b) dy \right\} db. \quad (7)$$

It can be noticed that: i) the term (5) is constant for any model specification p ; ii) (6) can be simplified to $-\int h_0(b) \log h(b) db$ which does not involve the conditional sampling model $f(y|b; \beta)$ and its minimization involves only the distribution of the random effects $h(b)$ which is not our focus; iii) the empirical version of the outer integral of (7) (see page 244 of [Linhart and Zucchini, 1986](#)) is

$$\tilde{\Delta}(f, f_0) = -\frac{1}{K} \sum_{i=1}^K \int f_{0i}(y_i|b_i) \log f_i(y_i; \hat{\beta}_i) dy_i, \quad (8)$$

where $f_i(y_i)$ is the pdf for data in cluster i , $\hat{\beta}_i$ is the cluster level estimate of $\beta + b_i$ and b_i is the random effect in cluster i . Going from (7) to (8) we use the fact that data in different clusters are independent under the generating

and the candidate model so that

$$\begin{aligned} \int \log f(y|b; \beta) f_0(y|b) dy &= \int \sum_i^K \log f_i(y_i; \beta + b_i) \prod_{j=1}^K f_{0j}(y_j|b_j) dy_j = \\ &= \sum_i^K \int \log f_i(y_i; \beta + b_i) f_{0i}(y_i|b_i) dy_i. \end{aligned}$$

An astute reader will have noticed that each term of the sum in (8) will be approximated up to a constant term by its counterpart from the cluster-specific Kulback-Leibler divergence:

$$\begin{aligned} \Delta_i(f, f_0) &= E_{f_0} \left\{ \log \frac{f_{0i}(Y_i|b_i)}{f_i(Y_i; \beta + b_i)} \right\} = \\ &= E_{f_0} \{ \log f_{0i}(Y_i|b_i) \} - E_{f_0} \left\{ \log f_i(Y_i; \hat{\beta}_i) \right\} \\ &= E_{f_0} \{ \log f_{0i}(Y_i|b_i) \} - \int f_{0i}(y_i|b_i) \log f_i(y_i; \hat{\beta}_i) dy_i. \end{aligned} \quad (9)$$

The cluster-specific AIC criterion computed using the data y_i and the MLE $\hat{\beta}_i$ is a consistent estimator of twice the second term in (9). Combining (7) and (9) leads to the mean AIC (*meanAIC*) model selection criterion:

$$\text{meanAIC} = \frac{1}{K} \sum_{i=1}^K \text{AIC}_i, \quad (10)$$

where $\text{AIC}_i = -2 \log f_i(y_i; \hat{\beta}_i) + 2(r+1)$ where $r+1$ is the dimension of β . Thus, meanAIC is simply the average of all K AICs obtained when fitting an ordinary GLM separately to each of the K clusters.

A heuristic justification of meanAIC is obtained by viewing the cluster-specific AICs as “scores” given to each candidate model. When H is a continuous distribution, with probability 1 the zero elements of the regression coefficient vectors β_i are the same in all clusters. Hence as $n_i \rightarrow \infty$ the best candidate model should receive the best score (lowest AIC) in each cluster. With finite samples, the best model may not get the lowest AIC uniformly in all K clusters, but after averaging over K clusters it is expected that meanAIC will identify the correct covariates; the simulation study reported in the next section confirms this intuition.

When averaging consistent estimators one may be tempted to give more weight to AIC_i 's derived from larger clusters. However, the terms that contribute to (10) are already implicitly weighted by the cluster size n_i , AIC_i already being a sum of n_i terms.

4. Simulation study

The meanAIC criterion was derived in the previous section as a potentially useful covariate selection tool when the cluster sizes n_i tend to infinity. The present section reports the results of a simulation study whose primary objective is to assess the performance of meanAIC as a covariate screening tool for finite values of n_i . A secondary objective is to compare its efficiency and robustness to that of mAIC that is computed assuming a GLMM with a random intercept. The latter choice is in line with our aim of establishing model selection criteria without having to spell out the random effects structure; under this constraint, the only specification that is common to all mixed effects submodels is the one with only a random intercept.

4.1. Study design

We generated samples with $K = 20$ independent clusters. Our primary objective suggests that we consider larger values of n_i . We ran simulations where n_i was fixed at 80 or 320 observations in all clusters; additional simulations where n_i varied from cluster to cluster within the same dataset yielded results similar to those reported below and are summarized in Appendix A. Our simulation design is similar to other designs where model selection criteria are investigated (e.g., Yu et al., 2013). We simulate two covariates, x_{ij1} iid Bernoulli(0.5) and x_{ij2} iid uniform(0, 1). The responses Y_{ij} are generated from a Poisson GLMM with log link and depend on x_{ij1} , a random intercept b_{0i} and a random coefficient b_{1i} through the following log conditional mean:

$$\log E[Y_{ij}; b_{0i}, b_{1i}] = 0.3 + b_{0i} + (\beta_1 + b_{1i})x_{ij1}, \quad (11)$$

for $1 \leq i \leq K$ and $1 \leq j \leq n_i$.

Our simulations consider two values for the fixed effects $\beta_1 \in \{0.2, 0.4\}$ and assumed the two random effects to be independently drawn from the same zero-mean normal distribution, but with possibly different variances, more precisely b_{ui} has variance σ_u^2 which can take values in the set $\{0.005, 0.15,$

0.3, 0.8, 1.5} for all $u \in \{0, 1\}$. We also simulate random effects from gamma and t distributions in order to investigate robustness of the criteria against asymmetric and fat-tailed distributions, respectively. These results are quite similar to those reported in Table 1 and are summarized in Appendix A.

For each sample generated, all four submodels were considered: i) the null model; ii) the true model with x_1 only; iii) the model with x_2 only, and iv) the model with both x_1 and x_2 . For each sample the following model selection criteria were computed: the proposed meanAIC and the mAIC obtained by fitting a Poisson GLMM with a random intercept to the entire sample.

The meanAIC is calculated using the output of the `glm` function from R applied to each cluster separately. Maximum likelihood fitting of the random intercept GLMM was implemented with the function `glmer` from the R package `lme4`. The calculation for meanAIC was approximately three times faster than the mAIC, e.g., computation for one sample required 0.05 second of CPU time for meanAIC and 0.17 second of CPU time for the mAIC on a Lenovo X230 tablet PC with Intel Core i7-3520M CPU at 2.90 GHz, 8 GB of RAM and running on the 64 bit version of Windows 7.

4.2. Results

Each simulation scenario was replicated 500 times and the proportion of correct covariate selection decisions are reported in Table 1. The simulations show that mAIC dominates meanAIC only when cluster-sizes are small and the random coefficient has a small variance. This is not surprising given that the mAIC criterion is computed assuming that only the intercept is random. Therefore, when averaging over the distribution of the random intercepts, mAIC pools all the cluster data and benefits from the resulting larger sample size, unlike meanAIC which relies on cluster-level data sizes that are not large enough to enable it to detect the small fixed effects. As soon as cluster-specific coefficients are not all small, e.g. the random coefficients have moderate variance, the accuracy of meanAIC dramatically increases. For example, when $n_i = 80$ and the random effect variance σ_1^2 increases from 0.005 to 0.15 we see that the percentage of correct decisions for meanAIC increases from about 12% to about 90% for different values of σ_0^2 . This trend is not replicated by mAIC that does not seem to benefit from larger values of σ_1^2 . When $n_i = 320$ both meanAIC and mAIC have a good performance, with meanAIC dominating mAIC for all values of σ_0^2 and σ_1^2 considered.

To see if the performance of meanAIC with $\sigma_1^2 = 0.005$ and $n_i = 80$

TABLE 1

Proportion of the 500 replications where each criterion picked the true model. Regression coefficient $\beta = 0.2$. Random effects follow a normal with mean 0 and variance σ_u^2 , $u = 0, 1$.

σ_0^2	σ_1^2	$n_i = 80 \ \forall i$		$n_i = 320 \ \forall i$	
		meanAIC	mAIC	meanAIC	mAIC
0.005	0.005	0.140	0.790	0.878	0.828
0.005	0.15	0.884	0.740	0.994	0.828
0.005	0.3	0.980	0.778	0.994	0.830
0.005	0.8	0.992	0.816	0.994	0.774
0.005	1.5	0.994	0.772	0.996	0.794
0.15	0.005	0.166	0.792	0.876	0.846
0.15	0.15	0.882	0.774	0.996	0.818
0.15	0.3	0.988	0.786	0.998	0.838
0.15	0.8	0.992	0.788	0.996	0.794
0.15	1.5	0.996	0.746	0.992	0.762
0.3	0.005	0.116	0.802	0.932	0.844
0.3	0.15	0.894	0.732	0.998	0.836
0.3	0.3	0.988	0.788	0.990	0.822
0.3	0.8	0.998	0.754	0.990	0.798
0.3	1.5	0.998	0.730	0.992	0.738

improves as the fixed effect size β gets larger, we replicate the simulation scenarios involving these values of n_i and σ_1^2 , but with a larger $\beta = 0.4$. The results are summarized in Table 2 and show that under these settings meanAIC outperforms mAIC.

5. Real Data Application

We illustrate the ability of meanAIC to identify those covariates that can have strong cluster-specific effects, but small marginal effects. These are typically covariates corresponding to important random effects variances that have modest fixed regression coefficients. The simulation studies performed in the previous section suggest that under this scenario mAIC and meanAIC are likely to select different models. The data come from the study of [Beauvais and Swaim \(2015\)](#) on alcohol use among young American Indians in U.S. schools. The individual observations are responses by students to a survey, and the clusters are the schools the students belong to. The sample we analyze consists of data from the 25 largest clusters with an average size of 170 students. The response variable is the number of times a student had

TABLE 2

Proportion of the 500 replications where each criterion picked the true model. Regression coefficient $\beta = 0.4$. Random effects follow a normal with mean 0 and variance σ_u^2 , $u = 0, 1$.

σ_0^2	σ_1^2	$n_i = 80 \ \forall i$	
		meanAIC	mAIC
0.005	0.005	0.866	0.806
0.15	0.005	0.910	0.856
0.3	0.005	0.928	0.854

more than 5 alcoholic drinks in less than two hours during the last two weeks.

All models considered will include four personal covariates (gender, age, whether the student likes school and whether the student is proud of him/herself), two covariates related to friends influence (whether some of the friends have ever been suspended from school and whether friends ask the student to get drunk some or a lot). The four candidate models differ due to presence/absence of two family-related covariates (family is a lot likely to stop the student from getting drunk, and the family members argue a lot). A more detailed description of the model covariates and their corresponding values is presented in Appendix B.

The models are fit using a Poisson GLMM with log link, with each school as a cluster. The meanAIC is obtained by fitting an ordinary Poisson GLM separately to each cluster data using the `glm` function in R. The mAIC is computed by fitting the marginal model using the `glmer` function from package `lme4` in R (Bates et al., 2015). Unlike meanAIC, to fit the GLMM needed for mAIC, one needs to specify a random effects structure. We consider two such structures: i) a random intercept only that will yield an mAIC value, denoted mAIC.RI, and ii) a random intercept and a random coefficient for the “family members argue a lot” covariate that will yield an mAIC value, denoted mAIC.RC. A summary of these GLMM fits is provided in Appendix B. Clearly, the random coefficient for “family members argue a lot” has a large variance and a small marginal effect. It is thus expected that meanAIC may identify more accurately than mAIC.RI the importance of this covariate for the model. The values of mAIC for the two GLMM and of meanAIC are reported in Table 3. Simulations showed that meanAIC is better than mAIC.RI at identifying the generating model covariates that had a random coefficient. Based on these results and the magnitude of the random coeffi-

TABLE 3

Model selection criteria for all four submodels of interest for the alcohol consumption study. *mAIC.RI* refers to the *mAIC* criterion obtained by fitting a GLMM with random intercept only, *mAIC.RC* denotes the *mAIC* of the model with a random intercept and a random coefficient in front of the covariate “family members argue a lot”, while *meanAIC* denotes the *meanAIC* criterion. The best value for each criterion appears in bold.

Family covariates in model	<i>mAIC.RI</i>	<i>mAIC.RC</i>	<i>meanAIC</i>
Both	7929.87	7898.47	298.47
“Lot likely to stop you” only	7928.37	7928.37	302.26
“Family members argue a lot” only	8166.20	8127.02	307.04
None	8166.06	8166.06	311.33

cient variance (this variance is highly significant according to the likelihood ratio test described in Section 6.3.2 of [Verbeke and Molenberghs \(2009\)](#)) reported in Appendix B, we believe that “family members argue a lot” should be part of the model. All criteria agree that the family is “likely to stop you from getting drunk” variable should be included in the model. It is worth pointing out that *mAIC* chooses different models depending on the random effects structure assumed, which can be confusing when there is no clear choice for the latter.

6. Discussion

In this paper we set out to develop a new variable selection criterion for GLMM that does not require a specification of the random effects structure. Furthermore, we wanted a criterion computable even when a two-stage estimation procedure is used to fit the model, which usually occurs when the cluster sizes are large enough to make impractical marginal likelihood inference.

We used an h-likelihood based theoretical justification to develop the *meanAIC* criterion. The implicit assumption is that cluster sizes are large. Simulations were performed under a number of possible combinations of cluster size (small and large), random coefficients variance (small, moderate and large), effect size (small and moderate) and different random effects distributions (normal, shifted gamma and t). We compared the ability of *meanAIC* and of *mAIC* to identify the true covariate structure. The proposed *meanAIC* clearly outperformed *mAIC* for all settings except the case where cluster size, random

TABLE 4

Proportion of the 500 replications where each criterion picked the true model. Regression coefficient $\beta = 0.2$. Random effects follow a shifted gamma distribution with shape parameter 4, mean 0 and variance σ_u^2 , $u = 0, 1$.

σ_0^2	σ_1^2	$n_i = 80 \forall i$		$n_i = 320 \forall i$	
		meanAIC	mAIC	meanAIC	mAIC
0.005	0.005	0.072	0.780	0.856	0.856
0.005	0.15	0.860	0.784	0.996	0.856
0.005	0.3	0.962	0.778	0.988	0.798
0.005	0.8	0.996	0.740	0.994	0.764
0.005	1.5	0.998	0.720	0.990	0.722
0.15	0.005	0.136	0.780	0.848	0.860
0.15	0.15	0.862	0.758	0.994	0.836
0.15	0.3	0.956	0.746	1.000	0.808
0.15	0.8	0.994	0.694	0.994	0.762
0.15	1.5	0.994	0.690	0.994	0.708
0.3	0.005	0.162	0.804	0.900	0.840
0.3	0.15	0.872	0.778	0.986	0.814
0.3	0.3	0.976	0.724	0.996	0.836
0.3	0.8	0.998	0.700	0.996	0.756
0.3	1.5	0.986	0.680	0.988	0.680

coefficient variance and fixed effect size were all simultaneously small. The application of these criteria to real data analysis further emphasized the importance of variable selection without specifying the random effects structure.

Model selection based on comparing all possible submodels is not practical when the number of potential covariates is large. In future work we would like to consider the interplay between meanAIC and regularization-based methods (e.g., [Ibrahim et al., 2011](#); [Fan and Li, 2012](#); [Lin et al., 2013](#)), where information criteria are used to set the value of the tuning parameter in the penalty term.

Acknowledgments

This work has been funded by Natural Sciences and Engineering Research Council of Canada individual discovery grants to each author.

Appendix A: Additional simulation results

TABLE 5

Proportion of the 500 replications where each criterion picked the true model. Regression coefficient $\beta = 0.2$. Random effects follow a t distribution with 3 degrees of freedom rescaled to have variance σ_u^2 , $u = 0, 1$.

σ_0^2	σ_1^2	$n_i = 80 \forall i$		$n_i = 320 \forall i$	
		meanAIC	mAIC	meanAIC	mAIC
0.005	0.005	0.122	0.786	0.838	0.822
0.005	0.15	0.752	0.794	0.996	0.850
0.005	0.3	0.928	0.736	0.990	0.828
0.005	0.8	0.984	0.750	0.996	0.748
0.005	1.5	0.988	0.682	0.998	0.696
0.15	0.005	0.116	0.820	0.886	0.820
0.15	0.15	0.752	0.736	0.998	0.822
0.15	0.3	0.910	0.744	0.996	0.794
0.15	0.8	0.990	0.736	0.994	0.782
0.15	1.5	0.990	0.716	0.992	0.762
0.3	0.005	0.154	0.776	0.866	0.824
0.3	0.15	0.774	0.726	0.986	0.852
0.3	0.3	0.944	0.748	1.000	0.806
0.3	0.8	0.992	0.760	0.994	0.772
0.3	1.5	0.998	0.868	0.994	0.718

Appendix B: Covariate description and summary of GLMM fits to the alcohol use data

Covariates

Female: 1 if student is female, 0 if not

Age: age of the student

DoNotLikeSchool: 1 if student does not like school, 0 otherwise

ProudSomeA lot: 1 if student is proud of him/herself a lot or some,
0 otherwise

Friends: 1 if some of the student's friends have been suspended from
school, 0 otherwise

FriendsAsk: 1 if student's friend ask student to get drunk some or a
lot, 0 otherwise

StopYou: 1 if student's family is a lot likely to stop student
from getting drunk, 0 otherwise

FamilyArgues: 1 if your family argues some or a lot, 0 otherwise

Random intercept model

Random effects:

TABLE 6

Proportion of the 500 replications where each criterion picked the true model. Regression coefficient $\beta = 0.2$. Random effects follow a normal with mean 0 and variance σ_u^2 , $u = 0, 1$.

σ_0^2	σ_1^2	$n_i \in \{40, 80, 160\}$	
		meanAIC	mAIC
0.005	0.005	0.156	0.790
0.005	0.15	0.932	0.760
0.005	0.3	0.984	0.800
0.005	0.8	0.994	0.796
0.005	1.5	0.996	0.708
0.15	0.005	0.194	0.798
0.15	0.15	0.926	0.746
0.15	0.3	0.988	0.776
0.15	0.8	0.992	0.758
0.15	1.5	0.990	0.752
0.3	0.005	0.208	0.810
0.3	0.15	0.928	0.754
0.3	0.3	0.988	0.774
0.3	0.8	1.000	0.772
0.3	1.5	0.988	0.766

```

Groups      Name      Variance Std.Dev.
SchoolID (Intercept) 0.2224  0.4716
Number of obs: 4232, groups: SchoolID, 25

```

Fixed effects:

```

              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.54128    0.31952  -7.953 1.81e-15
Female        -0.19667    0.04846  -4.058 4.94e-05
Age           0.10061    0.01941   5.185 2.17e-07
DoNotLikeSchool 0.41888    0.05743   7.294 3.02e-13
ProudSomeAlot -0.63376    0.05617 -11.284 < 2e-16
Friends       0.42097    0.06091   6.911 4.82e-12
FriendsAsk    1.42403    0.05958  23.901 < 2e-16
StopYou      -0.79153    0.05050 -15.674 < 2e-16
ArguesSomeAlot 0.03555    0.05007   0.710 0.478

```

Model with random intercept and random coefficient

Random effects:

```

Groups      Name      Variance Std.Dev. Corr

```



```
SchoolID (Intercept)      0.6534    0.8083
      ArguesSomeAlot 0.4340    0.6588    -0.85
Number of obs: 4232, groups: SchoolID, 25
```

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.57041	0.34870	-7.371	1.69e-13
Female	-0.20338	0.04861	-4.184	2.86e-05
Age	0.09203	0.01954	4.710	2.47e-06
DoNotLikeSchool	0.42672	0.05753	7.417	1.19e-13
ProudSomeAlot	-0.61726	0.05669	-10.889	< 2e-16
Friends	0.40415	0.06095	6.631	3.33e-11
FriendsAsk	1.42107	0.05945	23.904	< 2e-16
StopYou	-0.78243	0.05065	-15.449	< 2e-16
ArguesSomeAlot	0.21190	0.15486	1.368	0.171

References

- AKAIKE, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.*, **22** 203–217.
- BARRON, A., RISSANEN, J. and YU, B. (1998). The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory*, **44** 2743–2760.
- BATES, D., MÄCHLER, M., BOLKER, B. and WALKER, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67** 1–48.
- BEAUVAIS, F. and SWAIM, R. (2015). Drug use among young American Indians: Epidemiology and prediction, 1993-2006 and 2009-2013. Ann Arbor, MI: Inter-university Consortium for Political and Social Research. URL <http://doi.org/10.3886/ICPSR35062.v3>.
- BURNHAM, K. P. and ANDERSON, D. R. (2002). *Model Selection and Inference: A Practical Information-Theoretic Approach*. 2nd ed. Springer-Verlag New York Inc.
- CAVANAUGH, J. (1997). Unifying the derivations for the Akaike and corrected Akaike information criteria. *Statist. Probab. Lett.*, **33** 201–208.
- CHERVONEVA, I., IGLEWICZ, B. and HYSLOP, T. (2006). A general approach for two-stage analysis of multilevel clustered non-Gaussian data. *Biometrics*, **62** 752–759.

- CRAIU, R. V., DUCHESNE, T., FORTIN, D. and BAILLARGEON, S. (2011). Conditional logistic regression with longitudinal follow-up and individual-level random coefficients: A stable and efficient two-step estimation method. *J. Comput. Graph. Statist.*, **20** 767–784.
- CRAIU, R. V. and LEE, T. C. M. (2005). Model selection for the competing risks model with and without masking. *Technometrics*, **47** 457–467.
- DELATTRE, M., LAVIELLE, M. and POURSAT, M.-A. (2014). A note on BIC in mixed-effects models. *Electron. J. Statistics*, **8** 456–475.
- DONOHUE, M., OVERHOLSER, R., XU, R. and VAIDA, F. (2011). Conditional Akaike information under generalized linear and proportional hazards mixed models. *Biometrika*, **98** 685–700.
- FAN, Y. and LI, R. (2012). Variable selection in linear mixed effects models. *Ann. Statist.*, **40** 2043–2068.
- GREVEN, S. and KNEIB, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika*, **97** 773–789.
- HANSEN, M. and YU, B. (2003). Minimum description length model selection criteria for generalized linear models. In *Statistics and Science: A Festschrift for Terry Speed*. 145–163.
- HANSEN, M. H. and YU, B. (2001). Model selection and the principle of minimum description length. *J. Amer. Statist. Assoc.*, **96** 746–774.
- HURVICH, C. and TSAI, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76** 297–307.
- IBRAHIM, J. G., ZHU, H., GARCIA, R. I. and GUO, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics*, **67** 495–503.
- JIANG, J. (2007). *Linear and generalized linear mixed models and their applications*. Springer.
- JIANG, J., RAO, J., GU, Z. and NGUYEN, T. (2008). Fence methods for mixed model selection. *Ann. Statist.*, **36** 1669–1692.
- KORN, E. L. and WHITTEMORE, A. S. (1979). Methods for analyzing panel studies of acute health effects of air pollution. *Biometrics*, **35** 795–802.
- LEE, T. C. (2000). A minimum description length based image segmentation procedure, and its comparison with a cross-validation based segmentation procedure. *JASA*, **95** 259–270.
- LEE, T. C. M. (2001). An introduction to coding theory and the two-part minimum description length principle. *Int. Statist. Rev.*, **69** 169–183.
- LEE, Y. and NELDER, J. (1996). Hierarchical generalized linear models. *JRSS-B*, **58** 619–678.
- LI, L., YAO, F., CRAIU, R. V. and ZOU, J. (2014). Minimum description

- length principle for linear mixed effects models. *Statist. Sinica*, **24** 1161–1178.
- LIANG, H. and WU, H. (2008). A note on the conditional AIC for linear mixed-effects models. *Biometrika*, **95** 773–778.
- LIN, B., PANG, Z. and JIANG, J. (2013). Fixed and random effects selection by REML and pathwise coordinate optimization. *J. Comput. Graph. Statist.*, **22** 341–355.
- LINHART, H. and ZUCCHINI, W. (1986). *Model Selection*. Wiley.
- MOLENBERGHS, G., VERBEKE, G. and IDDI, S. (2011). Pseudo-likelihood methodology for partitioned large and complex samples. *Statist. Probab. Lett.*, **81** 892–901.
- MÜLLER, S., SCEALY, J. L. and WELSH, A. H. (2013). Model selection in linear mixed models. *Statist. Sci.*, **28** 135–167.
- NISHII, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, **12** 758–765.
- QUINLAN, J. and RIVEST, R. (1989). Inferring decision trees using the minimum description length principle. *Information and Computation*, **80** 227–248.
- RAO, C. R. and WU, Y. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika*, **76** 369–374.
- RENARD, D., MOLENBERGHS, G. and GEYS, H. (2004). A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics & Data Analysis*, **44** 649–667.
- RISSANEN, J. (1989). *Stochastic Complexity in Statistical Inquiry*. World Scientific Publishing Company, Singapore.
- SAEFKEN, B., KNEIB, T., VAN WAVEREN, C.-S. and GREVEN, S. (2014). A unifying approach to the estimation of the conditional Akaike information in generalized linear mixed models. *Electron. J. Statistics*, **8** 201–225.
- SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6** 461–464.
- SHANNON, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.*, **27** 379–423.
- SHIBATA, R. (2005). Approximate efficiency of a selection procedure for the number of regression variables. *Biometrika*, **92** 43–49.
- STIRATELLI, R., LAIRD, N. and WARE, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics*, **40** 961–971.
- VAIDA, F. and BLANCHARD, S. (2005). Conditional Akaike information for mixed effects models. *Biometrika*, **92** 351–370.

- VERBEKE, G. and MOLENBERGHS, G. (2009). *Linear Mixed Models for Longitudinal Data*. 2nd ed. Springer Science & Business Media, New York.
- YU, D., ZHANG, X. and YAU, K. K. W. (2013). Information based model selection criteria for generalized linear mixed models with unknown variance component parameters. *Journal of Multivariate Analysis*, **116** 245–262.